不完全な情報に基づく統計的検定

―― 対応のある場合の比率の差の検定 ――

山　口　　洋

〔抄　録〕

本稿では，出版物やインターネットで公表された集計値を 2 次利用して，対応のある場合の比率の差の検定を行う方法が提示された。この検定の p 値（有意確率）の正確な値を求めるには，2 変数のクロス表が必要だが，通常，公表されるのは 2 変数それぞれの度数分布だけである。しかし，本稿では，こうした場合でも p 値の最小値と最大値を求めることができ，うまくいけば，通常の検定を行ったときと同様の結論が得られることが明らかにされた。また，実際の集計値を用いた計算例が示された。その結果，本稿のアプローチは，クロス表が公表されておらず個票データが入手不可能・入手困難なときに有効であることが示された。

キーワード：対応のある場合の比率の差の検定，集計データ，2 次利用

1. 問題提起および本稿の目的とあらまし

書籍，論文，新聞，インターネットなどで公表されている無作為標本調査の集計値を 2 次利用するとき，そこに記載された集計値について独自に統計的検定をしたいと思うことがある。しかし，検定したい当の数値や，通常は公表されるごく基本的な情報（標本の大きさなど）の他に，通常は公表されないより詳細な情報を必要とする検定もある。この種の検定を普通の方法で行うには，しかるべき手続きをふんで個票データを入手するしかない。

しかし個票データが入手不可能または入手困難な場合も多く，古い時代の調査の個票データは，そもそも現存していないこともある。個票データが存在していれば，その所有者・管理者に検定を依頼することも考えられないではないが，所有者・管理者との間に特別な信頼関係が無いかぎり，結果を得るまでにはある程度の困難と時間を覚悟しなくてはならない。こうした場合，我々はこの種の検定を全面的にあきらめるべきなのか？

本稿の答えは「ノー」である。なぜなら，利用可能な限られた情報を使えば，厳密な p 値（有意確率）は求められなくても，ありうる p 値の区間，すなわち最大値と最小値を計算でき

ることがあるからだ。そしてこの p 値の存在範囲と予め設定した有意水準 α との関係によっては，通常の検定を行ったときと同様に「有意差あり」「有意差なし」といった結論を導くことができる。本稿の目的は，対応のある場合の比率の差の検定（安田・原 1982）について，こうしたアプローチの実例を示すことである。

対応のある場合の比率の差の検定は，次のような比率の差を検定するものである。すなわち，同じ対象者に同じ質問を繰り返し行ったときの各回の「はい」の比率の差，同じ対象者に行った異なるふたつの質問への「はい」の比率の差，無制限複数回答の質問における各選択肢の選択比率の差，夫とその妻における高学歴者（例えば四大卒以上）の比率の差，等々である。

後述するように，対応のある場合の比率の差の検定には，2 変数のクロス集計が必要になる。しかし，集計データを 2 次利用するケースでは，2 変数それぞれの度数分布しかわからない場合がほとんどである。この不完全な情報だけでは p 値の厳密な値を求めることはできない。しかし，クロス表（欠損値を除く）を作成したときの「周辺度数」を，2 変数それぞれの度数分布から特定できれば（できない場合は第 5〜6 節を参照），p 値がとりうる最大値と最小値を特定することができる。こうして特定された p 値の範囲が予め設定された有意水準（両側 5 %，1 % など）を完全に下回るならば，つまり p 値の最大値が有意水準を下回るならば，我々は「有意差あり」と判定して差し支えない。また，逆に p 値の範囲が有意水準を完全に上回るならば，つまり p 値の最小値が有意水準を上回るならば，我々は「有意差なし」と判定して差し支えない。p 値の範囲に有意水準が含まれるときは，いかなる判定も不可能であり「不明」とせざるをえない。しかし，このケースを除けば，我々は完全な情報を使って検定を行ったときと全く同様に「有意差あり」「有意差なし」といった判定を下すことができる。

以下，第 2 節で対応のある場合の比率の差の検定を定義する。第 3 節では p 値の最小値と最大値を求める方法を説明し，第 4 節で実用上の問題点と計算例を示す。第 5 節では第 3 節の方法が直接適用できない場合の対処法を述べ，第 6 節ではその計算例を示す。第 7 節では，以上をまとめて結論を述べる。また本稿の方法の根拠の証明は本稿末尾の補遺 1 と補遺 2 に示した。

2. 対応のある場合の比率の差の検定（二項検定）

対応のある場合の比率の差の主な検定方法として，マクネマー検定（安田・原 1982：261 頁）と，二項検定（同書：262 頁）がある。前者は後者の近似値を求める簡便法である。しかし現在では，パソコンの表計算ソフトを用いれば後者の方法を誰でも簡単に実施できる。また，どちらの方法も用いる情報は全く同じである。よって本稿はもっぱら二項検定について述べる。

二値（binary）の変数 X と Y を考える。具体例としては，無制限複数回答の質問（選択肢からいくつでも選択）におけるふたつの選択肢 X，Y について，各々「選択 =1，非選択 =0」とコード化するケース，また，夫の学歴（X）とその妻の学歴（Y）をそれぞれ「四大

卒以上 =1，それ以外 =0」とコーディングするケースなどをイメージしてほしい。そして，本稿では「X=1」の比率と「Y=1」の比率の差を検定する場合を考える。

この検定の帰無仮説は「母集団において X=1 の比率と Y=1 の比率は等しい」である。こう仮定したとき，実際のデータに生じた以上の比率の差が生じる確率，つまり p 値（有意確率）を求め，それがあらかじめ設定された有意水準 α（両側 5 %，片側 1 % といった）以下であれば帰無仮説を棄却し「有意差あり」と判定する。逆に p 値が有意水準 α を上回れば帰無仮説を棄却せず「有意差なし」と判定する。

この検定を行うには，表 1 のようなクロス集計表が必要である。各セルの記号は度数を表す。$n_{1.}$，$n_{0.}$，$n_{.1}$，$n_{.0}$，Nを特に「周辺度数」と呼ぶ。また，「対応のある場合の比率の差」とは表 1 の記号を使えば（$n_{1.}-n_{.1}$）/Nと表記することができる。

表 1　二値変数 X と Y のクロス表：
（ただし$n_{1.}>n_{.1}$，$n_{10}>n_{01}$）

	Y=1	Y=0	計
X=1	n_{11}	n_{10}	$n_{1.}$
X=0	n_{01}	n_{00}	$n_{0.}$
計	$n_{.1}$	$n_{.0}$	N

表内数値は度数

対応のある場合の比率の差の検定：片側 p 値を求めるための公式（ただし$n_{10}>n_{01}$）

$$p=\left(\frac{1}{2}\right)^{(n_{10}+n_{01})}\sum_{k=0}^{n_{01}}{}_{(n10+n01)}C_k \quad . \qquad \cdots\text{公式①}$$

記号　p：片側有意確率（両側有意確率は $2\times p$），${}_{(n_{10}+n_{01})}C_k$：$(n_{01}+n_{10})$ 個から k 個を選ぶ組合わせの数

片側 p 値を Microsoft Excel で求める数式

＝binomdist（n_{01}，$n_{10}+n_{01}$, 0.5, true）

表 1 の下には，このクロス表から比率の差の検定を行うための公式を記した（公式①）。公式①は X と Y の比率の差に直接関係するn_{10}とn_{01}のみに着目する。そして，$n_{10}+n_{01}$を一定と仮定したとき[(1)]，n_{10}とn_{01}の間にデータに現れた以上の差（データに現れた差を含む）が偶然現れる確率を計算するものである。なお表 1 では X=1 の比率（行側）が Y=1 の比率（列側）を上回るものと仮定しており，公式①はこの場合に対応している。また，公式①は片側 p 値を求めるものであり，両側 p 値はこの公式で求めた p 値を 2 倍して求める。

なお公式①の下には，この計算を表計算ソフト「Microsoft Excel」で行うための数式を示した。Excel のワークシートのセルに，この数式を入力して実行すればよい。ただし，数式中のセル度数を意味する部分には実際の数値を入力し，その他はイコール，（ ），コンマも含めてそのまま入力して実行する。たとえば$n_{10}+n_{01}$=123，n_{01}=45なら「＝binomdist（45, 123, 0.5, true）」とする。計算結果は 0.001113 である。

3. p 値の最小値・最大値を求めるアプローチ

前節でみたとおり，検定の p 値を求めるには表 1 のようなクロス表が必要である。しかし表 1 のうち周辺度数の部分，すなわち$n_{1.}$，$n_{0.}$，$n_{.1}$，$n_{.0}$，Nだけが分かれば，その情報を使って p 値の最大値と最小値を計算できる。つまり，ふたつの変数それぞれの度数分布から，クロス表の周辺度数が特定できれば（できない場合は後述），p 値の範囲が特定できる。この計算には次の性質を利用する。この性質の証明は本稿末尾の補遺 1 を参照してほしい。

「$n_{1.}-n_{.1}$（$=n_{10}-n_{01}$）が一定で，かつ$n_{1.}-n_{.1}\geq 2$（$\Leftrightarrow n_{10}-n_{01}\geq 2$）ならば，比率の差の検定の p 値は，$n_{10}+n_{01}$が最小のときに最小値を，$n_{10}+n_{01}$が最大のときに最大値をとる。」

なお$n_{10}-n_{01}=1$（$\Leftrightarrow n_{1.}-n_{.1}=1$）のとき，公式①で求めた片側 p 値は$n_{10}+n_{01}$によらず常に 0.5，両側 p 値は常に1となるから[(2)]，常に「有意差なし」という結論になる。

さて上の性質によれば，周辺度数を固定したとき，p 値が最小となるのはクロス表が表 2 のようになる場合である。表 2 の状態は正の最大関連と呼ばれる（安田 1969：39 頁）。このときの片側 p 値，すなわち片側 p 値の最小値を $\min(p)$ と表記すれば，それを求める公式は②のようになる。なお公式②は公式①に$n_{01}=0$，$n_{10}=n_{1.}-n_{.1}$を代入することで導かれる。なお，公式②の下には，この値を EXCEL で求めるときの数式を示した。

表 2　p 値が最小になる場合
（ただし$n_{1.}-n_{.1}\geq 2$）

	$Y=1$	$Y=0$
$X=1$		$n_{10}=n_{1.}-n_{.1}$
$X=0$	$n_{01}=0$	

片側 p 値の最小値：$\min(p)$ を求める公式

$$\min(p)=\left(\frac{1}{2}\right)^{(n_{1.}-n_{.1})}. \qquad \cdots\text{公式②}$$

EXCEL 数式 ＝(0.5)^($n_{1.}-n_{.1}$)

一方，p 値が最大となるのは，クロス表が表 3 または表 4 の状態にあるときで，これらの状態は負の最大関連と呼ばれる（安田 1969：40 頁）。表 3 は$n_{1.}+n_{.1}\leq N$のときの負の最大関連状態を，表 4 は$n_{1.}+n_{.1}\geq N$のときのそれを示す。ちなみに等号が成立するときは，$n_{11}=0$かつ$n_{00}=0$となり，この状態を特に正の完全関連と呼ぶ（安田 1969：39 頁）。表 3 および表 4 のときの片側 p 値，すなわち片側 p 値の最大値を $\max(p)$ と表記すれば，それは公式③および公式④により計算できる。公式③と④の下には EXCEL の数式を示した。ちなみに，正の完全関連状態のときは公式③と④のどちらを用いてもよい。つまりどちらを使っても同じ値が算出される。

表 3　p 値が最大になる場合
（$n_{1.}-n_{.1}\geq 2$ かつ$n_{1.}+n_{.1}\leq N$ のとき）

	$Y=1$	$Y=0$
$X=1$	$n_{11}=0$	$n_{10}=n_{1.}$
$X=0$	$n_{01}=n_{.1}$	

片側 p 値の最大値：$\max(p)$ を求める公式（その 1）

$$\max(p)=\left(\frac{1}{2}\right)^{(n_{1.}+n_{.1})}\sum_{k=0\,(n1.+n.1)}^{n_{.1}}C_k \quad . \qquad \cdots\text{公式③}$$

EXCEL 数式 ＝binomdist($n_{.1}$, $n_{1.}+n_{.1}$, 0.5, true)

表 4　p 値が最大になる場合
（$n_{1.}-n_{.1}\geq 2$かつ$n_{1.}+n_{.1}\geq N$のとき）

	$Y=1$	$Y=0$
$X=1$		$n_{10}=n_{.0}$
$X=0$	$n_{01}=n_{0.}$	$n_{00}=0$

片側 p 値の最大値：max(p) を求める公式（その 2）

$$\max(p)=\left(\frac{1}{2}\right)^{(n_{.0}+n_{0.})}\sum_{k=0}^{n_{0.}}{}_{(n_{.0}+n_{0.})}C_k \quad . \qquad \cdots公式④$$

EXCEL 数式 ＝binomdist($n_{0.}$, $n_{.0}+n_{0.}$, 0.5, true)

公式②～④で求めた min(p)，max(p)，および予め設定された有意水準 α の大小関係は次の 3 ケースに分類でき，それぞれ次のように結論づけることができる。

（1）$\max(p)\leq\alpha$　　（p 値の最大値が有意水準 α 以下）　　⇒「有意差あり」と結論

（2）$\alpha<\min(p)$　　（p 値の最小値が有意水準 α を超える）⇒「有意差なし」と結論

（3）$\min(p)\leq\alpha<\max(p)$　（p 値の最小値・最大値間に有意水準 α）⇒「不明」

（1）のとき，p 値がどんな値をとるにせよ，有意水準以下となることは確実だから，帰無仮説を棄却できて「有意差あり」と結論できる。同じく（2）のとき p 値が有意水準を上回ることは確実だから，帰無仮説を棄却せず「有意差なし」と結論できる。（3）のときは，有意水準 α のもとで，帰無仮説が棄却できるか否かが不明なので，どんな結論も下すことができない。つまり「不明」とするしかない。

検定結果のもうひとつの示し方として，有意水準 α を設定・明示せず，以下のように p 値のとりうる範囲を不等式で示す方法がある。

$$\min(p)\leq p\leq\max(p)$$

この不等式の min(p) と max(p) のところに実際の計算結果を代入して，$0.002\leq p\leq 0.015$ といった表記をすればよい。この表記をする場合，有意水準の設定とそれに基づく「5％水準で有意」「0.1% 水準で有意でない」「1％水準で有意であるか否かは不明」といった判定は読者に委ねられる。なお，上の不等式による表記は，通常の検定結果の表記で使われる不等式（両側 $p<.05$ といった）とは意味が違うので，注記等で説明が必要になるだろう。

4. 実用上の問題点と計算例　―― 表 1 の周辺度数が特定できる場合 ――

以上みてきたように，二値変数 X と Y のそれぞれの度数分布から表 1 の周辺度数，すなわち$n_{1.}$，$n_{0.}$，$n_{.1}$，$n_{.0}$，Nが特定できれば，その情報を使って p 値の最大値と最小値を計算でき，うまくいけば通常の検定と同様の判定を行いうる。これが本稿の方法の理論的な根幹部分である。本節以下では，実用上の問題点とその対処法について，実際の計算例を示しながら説明する。

前節の方法を実際に適用する際のネックは「欠損値」である。表 1 は「欠損値を除いたクロス表」なので，その周辺度数を特定するには，X と Y の双方が欠損値でないケースに限った

ときの度数分布が必要になる。度数分布データを欠損値の存在・不在，および存在の仕方で分類すると，以下の３種類になる。

（Ⅰ）欠損値が存在しない

（Ⅱ）「「Xが欠損値ならばYが欠損値」かつ「Yが欠損値ならばXが欠損値」」，が成立

（Ⅲ）「「Xが欠損値ならばYが欠損値」かつ「Yが欠損値ならばXが欠損値」」，が不成立

（Ⅰ）のときは，個々の変数の度数分布をそのまま表１の周辺度数と解釈できるから何の問題もない。一方，（Ⅱ）（Ⅲ）は欠損値が存在する場合だが，（Ⅱ）のときは容易に表１の周辺度数を特定できる。たとえば，無制限複数回答の質問での欠損値は，個々の選択肢（＝個々の変数）にではなく質問全体に割り当てられるから，（Ⅱ）のケースにあてはまる。やっかいなのは（Ⅲ）のケースである。複数の質問における「はい」の比率の差を検定したい場合の多くは，このカテゴリーに入る。（Ⅲ）の場合，変数X，Yそれぞれの度数分布がわかっても表１の周辺度数は特定できない。なぜなら一方の変数が正常値（1 か 0）で，他方の変数が欠損値であるケース（表１では削除される）が存在する可能性があるからだ。（Ⅲ）の場合の対処法は次節で述べることにしたい。

さて上で述べたように，無制限複数回答の質問は本稿のアプローチと非常に相性がよい。仮に欠損値が存在しても，必ず（Ⅱ）のケースに当てはまるからである。そこで本節では，無制限複数回答の質問の実例を使って，前節の方法の計算例を示すことにしよう。

2007 年に，NHK は日本全国の 16 歳以上の国民から無作為抽出(3)された回答者（有効数 2394 人，回収率 66.5%）に対して，様々な領域の「好きなもの」を配布回収法で調査した。そして，その結果を『日本人の好きなもの』という書籍にまとめている（NHK 放送文化研究所世論調査部編 2008）。調査項目の大半は無制限複数回答方式で質問されており，同書の巻末には各項目・各選択肢の選択比率が列記されている。ここから３項目ほどピックアップして，本稿の検定法の計算例を示そう。なお本節の本文中での計算はすべて，各質問項目に欠損値が存在しないものとして行われる。すなわち，各質問の「特になし＋無回答」(4) のカテゴリーを，双方非選択つまり「X=0 かつ Y=0」と同一視する。一方，「特になし＋無回答」を欠損値とみなした場合の結果も注（5）～（7）に記したので適宜参照してほしい。

この調査の問 17 では「するスポーツ」（見るスポーツでなく）で好きなものを 52 項目から無制限に複数回答させている。選択比率の第１位はボウリング（27.8%），第２位は野球（24.6%）であった。わずかながらボウリングが野球を上回る結果となっているが，この比率の差を有意水準両側５％で検定してみよう。最初に選択比率の高い「ボウリング」を X，低い「野球」を Y として，クロス表の周辺度数 $n_{1.}$，$n_{0.}$，$n_{.1}$，$n_{.0}$を求める（小数点以下は四捨五入）。

$n_{1.}=2394\times0.278=666$，$n_{0.}=2394-666=1728$

$n_{.1}=2394\times0.246=589$，$n_{.0}=2394-589=1805$　.

ここから両側 p 値の最小値を求めるには，片側 p 値を求める公式②に$n_{1.}$，$n_{1.}$を代入し，それを 2 倍すればよい。一方 p 値の最大値を求めるには，666＋589＝1255＜2394，つまり$n_{1.}+n_{.1}<N$だから，公式③に$n_{1.}$，$n_{.1}$を代入してそれを 2 倍する。以下のとおりである。

両側$\min(p)=2\times(1/2)^{(666-589)}=0.000$.

⇒ EXCEL 数式：「＝2 ＊ (0.5)＾(666－589)」→実行結果：「1.32E－23」

両側$\max(p)=2\times(1/2)^{(666+589)}\sum_{k=0}^{589}{}_{(666+589)}C_k=0.032$.

⇒ EXCEL 数式「＝2 ＊ (binomdist(589, 666＋589, 0.5, true))」→実行結果：「0.031886」

α＝0.05 だから，上の結果から 0.032＜0.05 となり，$\max(p)\leq\alpha$ だとわかる。つまり p 値の最大値が有意水準 α 以下だとわかる。よって，両側 5 % 水準で「有意差あり」と結論できる[(5)]。

同じ調査の問 25 では「好きな国・地域」を 52 項目から無制限に複数回答させた結果，第 1 位がオーストラリア（28.0%），第 2 位がイタリア（26.9%）であった。オーストラリアとイタリアの選択比率に差があると言えるか否かを有意水準両側 5 % で検定してみる。

$n_{1.}$＝2394×0.280＝670，$n_{0.}$＝2394－670＝1724,

$n_{.1}$＝2394×0.269＝644，$n_{.0}$＝2394－644＝1750　.

両側$\min(p)=2\times(1/2)^{(670-644)}=0.000$.

⇒ EXCEL 数式「＝2 ＊ (0.5)＾(670－644)」→実行結果：「2.98E－08」

670＋644＝1314＜2394，すなわち$n_{1.}+n_{.1}<N$なので公式③を用いて，

両側$\max(p)=2\times(1/2)^{(670+644)}\sum_{k=0}^{644}{}_{(670+644)}C_k=0.490$.

⇒ EXCEL 数式「＝2 ＊ (binomdist(644, 670＋644, 0.5, true))」→実行結果「0.490415」

α＝0.05 より，$\min(p)\leq\alpha<\max(p)$ である。よって両側 5 % 水準で，オーストラリアの選択比率とイタリアの選択比率に差があると言えるか否かは「不明」である[(6)]。

さらに，同じ調査の問 48 では，1 月～12 月までの中で好きな月を無制限に複数回答させている。選択比率のトップは 4 月（45.3%）で，わずかの差で 5 月（45.1%）が続く。以下，4 月と 5 月の選択比率に差があると言えるか否かを有意水準両側 5 % で検定してみる。

$n_{1.}$＝2394×0.453＝1084，$n_{0.}$＝2394－1084＝1310,

$n_{.1}$＝2394×0.451＝1080，$n_{.0}$＝2394－1080＝1314　.

両側$\min(p)=2\times(1/2)^{(1084-1080)}=0.125$.

⇒ EXCEL 数式「＝2 ＊ (0.5)＾(1084－1080)」→実行結果：「0.125」

1084＋1080＝2164＜2394，すなわち$n_{1.}+n_{.1}<N$なので公式③を用いて，

両側$\max(p)=2\times(1/2)^{(1084+1080)}\sum_{k=0}^{1080}{}_{(1084+1080)}C_k=0.949$.

⇒EXCEL 数式「＝2＊(binomdist(1080, 1084＋1080, 0.5, true))」→実行結果「0.948582」

α=0.05 より，$\alpha<\min(p)$ となり，p 値の最小値が有意水準 α を超えている。よって，「有意差なし」と結論できる。すなわち，4 月と 5 月の選択比率に差があるとは言えない(7)。

5. 表 1 の周辺度数が特定できないときの方法

本節では，前節で述べた（Ⅲ）の場合の対処法について述べる。（Ⅲ）の場合，欠損値の数が相対的に多ければ本稿のアプローチは断念せざるをえない。しかし，欠損値の度数が相対的に少なければ，後述する方法を使って p 値の最大値・最小値が計算できる。「欠損値が相対的に少ない」ということの厳密な定義は後ほど下の表 5 を使って示そう。

表 5 は欠損値を含む 3×3 のクロス表である。$X=m$ は変数 X が欠損値，$Y=m$ は変数 Y が欠損値であることを示す。セル内の記号は度数を表す。表 5 の $X=m$ の行と $Y=m$ の列を削除したのが表 1 である。本節では，表 5 の周辺度数（$n'_{1.}$，$n'_{0.}$，$n'_{m.}$，$n'_{.1}$，$n'_{.0}$，$n'_{.m}$）すなわち，変数 X と Y それぞれの度数分布だけがわかっている状況を想定する。

表 5　欠損値を含むクロス表
（ただし $n'_{1.}>n'_{.1}$）

	$Y=1$	$Y=0$	$Y=m$	計
$X=1$	n_{11}	n_{10}	n_{1m}	$n'_{1.}$
$X=0$	n_{01}	n_{00}	n_{0m}	$n'_{0.}$
$X=m$	n_{m1}	n_{m0}	n_{mm}	$n'_{m.}$
計	$n'_{.1}$	$n'_{.0}$	$n'_{.m}$	N'

表内数値は度数

さて，欠損値の度数が「相対的に少ない」とは，次の 2 個の前提が満たされる場合をさす。本節で述べる方法は，これらの前提を満たす場合にのみ適用可能である。

前提 1：$n'_{m.}<n'_{.1}$かつ$n'_{m.}<n'_{.0}$かつ$n'_{.m}<n'_{1.}$かつ$n'_{.m}<n'_{0.}$．

前提 2：$n'_{.m}<n'_{1.}-n'_{.1}$ ．

前提 1 が満たされないと，一方の変数が欠損値であるケースを削除したとき，他方の変数の正常値のいずれかが完全に失われ，表 1 のような 2×2 のクロス表を作成できなくなる可能性がある。前提 2 が満たされないと，欠損値を除く 2×2 のクロス表を作成したとき，表 5 での比率の差が消失または逆転する可能性がある。

では，これらの前提が満たされる場合の方法を説明する。本節の作業は，表 5 の周辺度数を固定したとき，ありうる最小の p 値と最大の p 値を求めることである。表 5 の周辺度数を固定しても，欠損値の分布（n_{1m}，n_{0m}，n_{m1}，n_{m0}，n_{mm}）に応じて，表 1 の周辺度数（$n_{1.}$，$n_{0.}$，$n_{.1}$，$n_{.0}$，N）の分布は変わり，その変化に応じて公式②～④で計算される p 値の最小値，最大値も変動する。したがって本節の作業をより正確に表現すれば，このように変動する p 値の最小値の最小値：$\min\{\min(p)\}$，最大値の最大値：$\max\{\max(p)\}$ を求めることである。なお，この考え方では，欠損値を除く 2×2 クロス表での比率の差それ自体が幅をもった値とな

る。後述するようにp値の最小値は比率の差の最大値に，p値の最大値は比率の差の最小値に対応する。よって分析結果を提示する際には，この比率の差の範囲も示した方がよい。

表5の周辺度数を固定したとき，表1において比率の差$(n_{1.}-n_{.1})/N$が最大となるのは，「X=1」のケースに「Y=欠損値」のケースが一切含まれず，「Y=1」のケースに「X=欠損値」のケースのすべてが含まれる場合，つまり表5で$n_{1m}=0$，$n_{0m}=n'_{.m}$，$n_{m1}=n'_{m.}$，$n_{m0}=0$，$n_{mm}=0$となる場合である(8)。このとき公式②で求めるmin(p)は最小となる（証明は補遺2を参照）。したがってp値の最小値とそれに対応する比率の差の最大値の求め方は以下のようになる。

表1の周辺度数が特定できない場合の，p値の最小値と比率の差の最大値の求め方

$n_{1.}=n'_{1.}$．　→公式②に代入

$n_{0.}=n'_{0.}-n'_{.m}$　．

$n_{.1}=n'_{.1}-n'_{m.}$．→公式②に代入

$n_{.0}=n'_{.0}$．

$N=n_{1.}+n_{0.}=n'_{1.}+n'_{0.}-n'_{.m}$　．

比率の差の最大値：$(n_{1.}-n_{.1})/N=\{n'_{1.}-(n'_{.1}-n'_{m.})\}/(n'_{1.}+n'_{0.}-n'_{.m})$　．

前提1：$n'_{m.}<n'_{.1}$かつ$n'_{m.}<n'_{.0}$かつ$n'_{.m}<n'_{1.}$かつ$n'_{.m}<n'_{0.}$．

前提2：$n'_{.m}<n'_{1.}-n'_{.1}$．

一方，表1において比率の差$(n_{1.}-n_{.1})/N$が最小となるのは，「X=1」のケースに「Y=欠損値」のケースのすべてが含まれ，「Y=1」のケースに「X= 欠損値」のケースが一切含まれない場合，すなわち，表5で$n_{1m}=n'_{.m}$，$n_{0m}=0$，$n_{m1}=0$，$n_{m0}=n'_{m.}$，$n_{mm}=0$となる場合である(9)。このとき公式③④で求めるmax(p)は最大となる（証明は補遺2）。よってp値の最大値の求め方は以下のようになる。

表1の周辺度数が特定できない場合の，p値の最大値と比率の差の最小値の求め方

公式③は，$n_{1.}+n_{.1}\leq N$のときに，公式④は$n_{1.}+n_{.1}\geq N$のときに用いる。

$n_{1.}=n'_{1.}-n'_{.m}$．　→公式③に代入

$n_{0.}=n'_{0.}$　．　→公式④に代入

$n_{.1}=n'_{.1}$　．　→公式③に代入

$n_{.0}=n'_{.0}-n'_{m.}$．→公式④に代入

$N=n_{1.}+n_{0.}=n'_{1.}+n'_{0.}-n'_{.m}$．

比率の差の最小値：$(n_{1.}-n_{.1})/N=\{(n'_{1.}-n'_{.m})-n'_{.1}\}/(n'_{1.}+n'_{0.}-n'_{.m})$　．

前提1：$n'_{m.}<n'_{.1}$かつ$n'_{m.}<n'_{.0}$かつ$n'_{.m}<n'_{1.}$かつ$n'_{.m}<n'_{0.}$　．

前提 2：$n'_{.m} < n'_{1.} - n'_{.1}$　.

6. 計　算　例　―― 表 1 の周辺度数が特定できない場合 ――

以下では，前節の方法の計算例を示す。1995 年の職業威信調査[10]では，全国の 20 歳～69 歳の有権者から無作為抽出[11]された回答者（有効数 1214 人，回収率 72.5%）に対して個別訪問面接を行い，56 種類の職業の威信を，それぞれ「最も高い」～「最も低い」の 5 段階で評価させている（原編 2000，都築編 1998）。この調査の基礎集計表によれば，「最も高い」という回答の比率は第 1 位が「医師」（64.3%，781 人），第 2 位が「大会社の社長」（57.0%，692 人）であった（1995 年 SSM 調査研究会編 1997）。また「DK, NA」の比率は，医師の質問で 2.6%（32 人），大会社の社長の質問で 2.7%（33 人）だった。ここでは「DK, NA」を欠損値とみなし，前節の方法を使って「医師＝最も高い（$X=1$）」の比率が「大会社の社長＝最も高い（$Y=1$）」の比率を上回ると言えるかどうかを，有意水準片側 5 % で検定してみる。この例において表 5 の周辺度数は以下のようになる。

$n'_{1.}=781$，$n'_{.m}=32$，$n'_{0.}=1214-781-32=401$
$n'_{.1}=692$，$n'_{.m}=33$，$n'_{.0}=1214-692-33=489$.

よって前提 1（$n'_{.m} < n'_{.1}$かつ$n'_{.m} < n'_{.0}$かつ$n'_{.m} < n'_{1.}$かつ$n'_{.m} < n'_{0.}$）は満たされている。また $33 < 781-692=89$ なので前提 2（$n'_{.m} < n'_{1.} - n'_{.1}$）も満たす。したがって，この事例には前節の方法が適用可能である。そこでまず片側 p 値の最小値，およびそれに対応する比率の差の最大値を求める。

$n_{1.}=n'_{1.}=781$.
$n_{0.}=n'_{0.}-n'_{.m}=401-33=368$.
$n_{.1}=n'_{.1}-n'_{.m}=692-32=660$.
$n_{.0}=n'_{.0}=489$.
$N=n_{1.}+n_{0.}=781+368=1149$　.
比率の差の最大値：$(n_{1.}-n_{.1})/N=(781-660)/1149=0.105$　(10.5%)　.
片側$\min(p)=(1/2)^{(781-660)}=0.000$.
⇒ EXCEL 数式：「＝(0.5)^(781－660)」→実行結果：「3.76E－37」

次に，片側 p 値の最大値，およびそれに対応する比率の差の最小値を求める。

$n_{1.}=n'_{1.}-n'_{.m}=781-33=748$.

$n_{0.}=n'_{0.}=401$.

$n_{.1}=n'_{.1}=692$.

$n_{.0}=n'_{.0}-n'_{m.}=489-32=457$.

$N=n_{1.}+n_{0.}=748+401=1149$.

比率の差の最小値：$(n_{1.}-n_{.1})/N=(748-692)/1149=0.049$ （4.9%）.

$n_{1.}+n_{.1}=781+692=1473>1149=N$なので，公式④を用いる。

片側 $\max(p)=(1/2)^{(457+401)}\sum_{k=0\,(457+401)}^{401}C_k=0.030$.

⇒ EXCEL 数式「＝binomdist(401, 457＋401, 0.5, true)」→実行結果「0.030182」

以上の結果，$0.000 \leq p \leq 0.030 < 0.05=\alpha$ となるから，$\max(p) \leq \alpha$ だとわかる。よって片側５％水準で「有意差あり」と言うことができる。つまり，母集団においても「医師＝最も高い」とする人の比率は，「大会社の社長＝最も高い」とする人の比率を上回っている，と言うことができる。また，欠損値を除くクロス表を作成したときの両者の比率の差は 4.9% 以上，10.5% 以下となり，上記の p 値の幅はこの比率の差の幅に対応するものである。つまり $\min(p)=0.000$ という結果は，比率の差が 10.5%（最大）であったときの最小の p 値を意味し，$\max(p)=0.030$ という結果は，比率の差が 4.9%（最小）であったときの最大の p 値を意味している。

7. 要約と結論

対応のある場合の比率の差の検定の正確な p 値を求めるには，比率の差を比較する 2 変数のクロス表（第 2 節参照）が必要だが，集計データを 2 次利用する場合，個々の変数の度数分布しかわからないことが多い。本稿はこのような場合に有効な，p 値の最小値・最大値を求めるアプローチを提案し（第 3 節，第 5 節），実際の集計データを用いて計算例を示した（第 4 節，第 6 節）。対応のある場合の比率の差の検定が適用されるケースの中には，本稿のアプローチと比較的相性が良いもの（第 4 節）と，そうでないもの（第 6 節）がある。相性が良いのは，2 変数とも正常値（欠損値でない）をとるケースに限ったときの度数分布がわかる場合である。このとき，第 3 節の方法を直接用いることができる。相性が良くないのは，欠損値が変数ごとに別々にカウントされていて，2 変数とも正常値をとるケースに限ったときの度数分布がわからない場合である。しかしこの場合でも，欠損値の数が相対的に少なければ（定義は第 5 節参照），第 5 節の手続きを経た後，第 3 節の方法を適用することによって，p 値の最小値と最大値を求めることができる。

ちなみに，対応のある場合の比率の差の検定だけでなく，対応のある場合の平均値の差の検

定でも，p 値の最小値・最大値を求める同様のアプローチを提示できる。2 変数間の相関係数が公表されていなければ，この検定の p 値は求められないが，2 変数それぞれの平均値と標準偏差が公表されていれば，p 値の最小値と最大値を求めることは可能である。具体的な方法の提示は今後の課題としたい。

最後に，本稿の意義を再確認して締めくくりたい。言うまでもなく本稿のアプローチは，2 変数のクロス表が公表されている状況，または個票データが手元にあってクロス表が作成可能な状況では不要である。このときは，公式①を使って比率の差の p 値の正確な値を求めればよい。本稿のアプローチが最も必要とされるのは，クロス表が公表されておらず，かつ個票データが現存していない場合または現存していても入手不可能な場合である。このとき，表 1 のクロス表の作成は不可能なので，本稿のアプローチが検定のための唯一残された手段となる。

また，個票データが入手不可能ではないが，入手のための手続きや分析の手間を考慮すると，これを躊躇してしまう場合にも，本稿のアプローチは有効であろう。単にいくつかの比率の差を検定したいだけなら，本稿のアプローチで「有意差あり」「有意差なし」といった結論が得られればそれで十分なのであり，種々のやっかいな手続きを踏んでまで個票データを入手・分析する必要はない（もちろん公表された集計値自体が信頼に値するものであることが前提だが)。個票データの入手・分析を検討するのは，本稿のアプローチで結論が「不明」となったとき，もしくは欠損値が多いために本稿のアプローチが適用不能（第 5 節参照）となったときで遅くない。

また，本稿が疑問を呈したかったのは「正確な数値が求められないなら，一切をあきらめるべきである」という「全か無か」の考え方である。おそらく，こうした考え方は怠惰からではなく，むしろ「可能な限り正確な数値を追い求めなくてはならない」という観念からくるものである。しかし，そうしたいささか強迫的な観念から少し自由になってみると，我々は限られた情報からでも，かなりのことを知りうることに気づく。対応のある場合の比率の差の検定は，既存の個票データをしかるべき手続きで入手するか，自ら調査を行って個票データを収集するかして行うのが正しい方法であることは言うまでもない。しかし，それが不可能または困難だったとしても，すべてをあきらめてしまう必要はないのである。

補遺 1　表 1 の周辺度数を固定したときの p 値の最大・最小値に関する命題の証明

「表 1 で $n_{10}-n_{01}$ が一定（ただし $n_{10}-n_{01}\geq 2$）ならば，対応のある場合の比率の差の検定の p 値は，$n_{10}+n_{01}$ が最小のときに最小値を，$n_{10}+n_{01}$ が最大のときに最大値をとる」という命題を証明する。これを証明するには，下の表 6 における p 値を，その右の表 7 における p 値が必ず上回ることを証明すればよい（表 6 と表 7 における $n_{10}-n_{01}$ は等しい)。これが証明できれば，$n_{10}+n_{01}$ が最大のときに p 値が最大，$n_{10}+n_{01}$ が最小のときに p 値が最小となることは自明である。よって表 6 および表 7 の a と b が，$a-b\geq 2$ かつ $b\geq 0$ を満たす任意の整数である

とき，以下の不等式（ｉ）が成立することを証明すればよい。

表6　クロス表（$a-b\geq 2$かつ$b\geq 0$）

	$n_{10}=a$
$n_{01}=b$	

表7　クロス表（同左）

	$n_{10}=a+1$
$n_{01}=b+1$	

$$\left(\frac{1}{2}\right)^{a+b+2}\sum_{k=0}^{b+1}{}_{a+b+2}C_k-\left(\frac{1}{2}\right)^{a+b}\sum_{k=0}^{b}{}_{a+b}C_k=\left(\frac{1}{2}\right)^{a+b+2}\left\{\sum_{k=0}^{b+1}{}_{a+b+2}C_k-4\sum_{k=0}^{b}{}_{a+b}C_k\right\}>0.$$

…不等式（ｉ）

式（ｉ）の中カッコ内の第1項は，組み合わせの数の公式，

$$\sum_{r=0}^{m}{}_{n+1}C_r=2\sum_{r=0}^{m-1}{}_nC_r+{}_nC_m \qquad \cdots(\text{ii})$$

（森口・宇田川・一松 1987：11頁6行目）

の左辺に$r=k$，$m=b+1$，$n=a+b+1$を代入したものだから，次のように変形できる。

$$\sum_{k=0}^{b+1}{}_{a+b+2}C_k=2\sum_{k=0}^{b}{}_{a+b+1}C_k+{}_{a+b+1}C_{b+1}=2\sum_{k=0}^{b+1}{}_{a+b+1}C_k-{}_{a+b+1}C_{b+1} \quad . \qquad \cdots(\text{iii})$$

一方，公式（ii）に$r=k$，$m=b+1$，$n=a+b$を代入すると，以下のようになる。

$$\sum_{k=0}^{b+1}{}_{a+b+1}C_k=2\sum_{k=0}^{b}{}_{a+b}C_k+{}_{a+b}C_{b+1} \quad . \qquad \cdots(\text{iv})$$

（iv）を（iii）に代入し，組合せの数の公式${}_nC_r={}_{n-1}C_{r-1}+{}_{n-1}C_r$を使って変形すると，

$$\sum_{k=0}^{b+1}{}_{a+b+2}C_k=4\sum_{k=0}^{b}{}_{a+b}C_k+2{}_{a+b}C_{b+1}-{}_{a+b}C_{b+1}-{}_{a+b}C_b=4\sum_{k=0}^{b}{}_{a+b}C_k+{}_{a+b}C_{b+1}-{}_{a+b}C_b \quad .$$

よって，（ｉ）の中カッコ内は${}_{a+b}C_{b+1}-{}_{a+b}C_b$となる。ここで$2\leq a-b$より$b+2\leq a$だから，$b+1\leq\frac{1}{2}(a+b)$となるので，組合せの数の性質により必ず，${}_{a+b}C_{b+1}>{}_{a+b}C_b$となる。よって，（ｉ）の中カッコ内は正。$(1/2)^{a+b+2}$も正なので，不等式（ｉ）の成立が証明された。

補遺2　表5の周辺度数を固定したときのp値の最大値・最小値に関する命題の証明

表5　再掲（$n'_{1.}>n'_{.1}$）

	$Y=1$	$Y=0$	$Y=m$	計
$X=1$	n_{11}	n_{10}	n_{1m}	$n'_{1.}$
$X=0$	n_{01}	n_{00}	n_{0m}	$n'_{0.}$
$X=m$	n_{m1}	n_{m0}	n_{mm}	$n'_{m.}$
計	$n'_{.1}$	$n'_{.0}$	$n'_{.m}$	N'

表8

	$Y=1$	$Y=0$	$Y=m$
$X=1$		$n_{10}=n_{1.}-n_{.1}$	0
$X=0$	0		$n'_{.m}$
$X=m$	$n'_{m.}$	0	0

表9　（$n_{1.}+n_{.1}\leq N$のとき）

	$Y=1$	$Y=0$	$Y=m$
$X=1$	0	$n_{10}=n_{1.}$	$n'_{.m}$
$X=0$	$n_{01}=n_{.1}$		0
$X=m$	0	$n'_{m.}$	0

表10　（$n_{1.}+n_{.1}\geq N$のとき）

	$Y=1$	$Y=0$	$Y=m$
$X=1$		$n_{10}=n_{.0}$	$n'_{.m}$
$X=0$	$n_{01}=n_{0.}$	0	0
$X=m$	0	$n'_{m.}$	0

表5の周辺度数を固定すると，比率の差の検定のp値は表8のときに最小値を，同じく表9または表10のときに最大値をとることを証明する。なお，セル度数の記号法は表1と表5に準拠する。またセル度数に関する仮定は表8～表10まで共通に$n_{10}-n_{01}\geq 2$とし，表9では$n_{1\cdot}+n_{\cdot 1}\leq N$，表10では$n_{1\cdot}+n_{\cdot 1}\geq N$という仮定を加える。

まず，表5の周辺度数を固定すると，表8のときのp値が，ありうるp値の最小値であることを示す。それには，公式②で求める$\min(p)$が表8で最も小さくなることを示せばよい。公式②より$\min(p)$は$n_{1\cdot}-n_{\cdot 1}$が最大のときに最小値をとる。それは$n_{1\cdot}$が最大かつ$n_{\cdot 1}$が最小のときである。表5の周辺度数を固定すると，$n_{1\cdot}$が最大となるのは$n_{1m}=0$のときで，$n_{\cdot 1}$が最小となるのは$n_{m1}=n'_{m\cdot}$のときである。よって，表5の周辺度数を固定したときの最小のp値は，表8のときのp値である。証明終わり。

次に，表5の周辺度数を固定すると，表9（$n_{1\cdot}+n_{\cdot 1}\leq N$のとき）または表10（$n_{1\cdot}+n_{\cdot 1}\geq N$のとき）での$p$値が，$p$値の最大値であることを示す。それには，公式③で求める$\max(p)$が表9で，公式④で求める$\max(p)$が表10で，それぞれ最大となることを示せばよい。

下の性質1と2（その証明は後述）から，n_{10}が最小かつn_{01}が最大のとき，$\max(p)$は最大値をとるとわかる。公式③が適用されるとき$n_{10}=n_{1\cdot}$，$n_{01}=n_{\cdot 1}$であるから，$n_{1\cdot}$が最小かつ$n_{\cdot 1}$が最大のとき$\max(p)$は最大となる。表5の周辺度数を固定したとき，$n_{1\cdot}$が最小となるのは$n_{1m}=n'_{\cdot m}$のときであり，$n_{\cdot 1}$が最大となるのは$n_{m1}=0$のときである。よって，表5の周辺度数を固定したときの最大のp値は，$n_{1\cdot}+n_{\cdot 1}\leq N$の場合，表9のときの$p$値である。一方，公式④が適用される状況において$n_{10}=n_{\cdot 0}$，$n_{01}=n_{0\cdot}$であるから，$n_{\cdot 0}$が最小かつ$n_{0\cdot}$が最大のとき$\max(p)$は最大となる。$n_{\cdot 0}$が最小となるのは$n_{m0}=n'_{m\cdot}$のときであり，$n_{0\cdot}$が最大となるのは$n_{0\mathrm{m}}=0$のときである。よって，表5の周辺度数を固定したときの最大のp値は，$n_{1\cdot}+n_{\cdot 1}\geq N$の場合には，表10のときの$p$値である。証明終わり。

性質1　表1のn_{01}が一定なら，n_{10}が大きければ大きいほどp値は小さくなる。
性質2　表1のn_{10}が一定なら，n_{01}が大きければ大きいほどp値は大きくなる。

表6　再掲（$a-b\geq 2$かつ$b\geq 0$）

	$n_{10}=a$
$n_{01}=b$	

表11　（同左）

	$n_{10}=a+1$
$n_{01}=b$	

表12　（同左）

	$n_{10}=a$
$n_{01}=b+1$	

性質1の証明。これを示すには表6でのp値よりも表11でのp値が必ず小さくなることを示せばよい。表6のp値から表11のp値を引いた値は以下のように表現される。

$$\left(\frac{1}{2}\right)^{a+b}\sum_{k=0}^{b}{}_{a+b}C_k-\left(\frac{1}{2}\right)^{a+b+1}\sum_{k=0}^{b}{}_{a+b+1}C_k=\left(\frac{1}{2}\right)^{a+b+1}\left\{2\sum_{k=0}^{b}{}_{a+b}C_k-\sum_{k=0}^{b}{}_{a+b+1}C_k\right\}.\cdots(\mathrm{v})$$

一方，補遺1の公式（ii）を変形すると，$\sum_{r=0}^{m}{}_{n+1}C_r=2\sum_{r=0}^{m}{}_nC_r-{}_nC_m$となり，ここに

$r=k$，$m=b$，$n=a+b$ を代入すると，$\sum_{k=0}^{b}{}_{a+b+1}C_k=2\sum_{k=0}^{b}{}_{a+b}C_k-{}_{a+b}C_b$ となるので，（v）の中カッコ内は${}_{a+b}C_b$となり，これは必ず正。また，$(1/2)^{a+b+1}$も正だから（v）全体は必ず正。よって，表6のp値より表11のp値は必ず小さくなる。証明終わり。

性質2の証明。これを示すには，表12でのp値が表6でのp値よりも大きくなることを証明すればよい。表12のp値から表6のp値を引いた値は以下のように表現される。

$$\left(\frac{1}{2}\right)^{a+b+1}\sum_{k=0}^{b+1}{}_{a+b+1}C_k-\left(\frac{1}{2}\right)^{a+b}\sum_{k=0}^{b}{}_{a+b}C_k=\left(\frac{1}{2}\right)^{a+b+1}\left\{\sum_{k=0}^{b+1}{}_{a+b+1}C_k-2\sum_{k=0}^{b}{}_{a+b}C_k\right\}.\cdots\text{（vi）}$$

補遺1の（iv）より，（vi）の中カッコ内は${}_{a+b}C_{b+1}$となり，これは必ず正。$(1/2)^{a+b+1}$も正なので（vi）全体も正。よって，表6のp値より表12のp値は必ず大きくなる。証明終わり。

注

（1）安田・原（1982：263頁）は，この仮定が「いつも妥当するとはかぎらないだろう」と述べるが，具体的にどんな場合に妥当しないのかの説明は無い。

（2）このことの証明は以下のとおり。補遺1の表6で$a=b+1$のとき，比率の差のp値は常に0.5となることを証明する。表6で$a=b+1$のとき片側p値を求める式は次のとおり。$p=(1/2)^{(2b+1)}\sum_{k=0}^{b}{}_{2b+1}C_k\cdots$(vii)。一方，二項定理（$(a+b)^n=a^rb^{n-r}\sum_{r=0}^{n}{}_nC_r$），および二項係数の性質（${}_nC_r={}_nC_{n-r}$）から，次の等式が成り立つ。$2^{2b+1}=(1+1)^{2b+1}=\sum_{k=0}^{2b+1}{}_{2b+1}C_k=\sum_{k=0}^{b}{}_{2b+1}C_k+\sum_{k=b+1}^{2b+1}{}_{2b+1}C_k=2\sum_{k=0}^{b}{}_{2b+1}C_k$．したがって以下の等式が成り立つ。$\sum_{k=0}^{b}{}_{2b+1}C_k=2^{2b+1}\times\frac{1}{2}=2^{2b}$．この結果を(vii)に代入すると，$p=(1/2)^{2b+1}2^{2b}=1/2=0.5$となるので，補遺1の表6で$a=b+1$のとき，比率の差の$p$値は常に0.5となることが証明された。

（3）住民基本台帳から層化無作為2段抽出により3600人（12人×300地点）を抽出している。調査有効数はそのうち2394人（回収率66.5%）である。

（4）原典では「特になし」「無回答」それぞれの度数は不明で，「特になし＋無回答」の度数しかわからない（NHK放送文化研究所世論調査部編 2008）。しかし，「特になし」は「$X=0$ かつ $Y=0$」と問題なく同一視できるのに対して，「無回答」には「$X=$ 欠損値かつ $Y=$欠損値」と同一視すべきケースも含まれている可能性がある。そこで本稿では，「特になし＋無回答」を「$X=0$ かつ $Y=0$」と同一視した場合について本文で述べ，「$X=$欠損値かつ $Y=$欠損値」と同一視した場合について，以下の注記で述べることにした。

（5）この問17において「特になし＋無回答（17.5%）」を欠損値とみなすと，2394×0.175＝419人が欠損値である。よってボウリングの非選択者数は$n_{0.}$＝2394－666－419＝1309，野球の非選択者数は$n_{.0}$＝2394－589－419＝1386，N＝2394－419＝1975となる。しかし選択者数$n_{1.}$，$n_{.1}$に変化は無いので，公式②で求めるp値の最小値は本文中の値と同じになる。また，666＋589＝1255＜1975（$n_{1.}+n_{.1}<N$）なのでp値の最大値の計算には公式③を用いるから，この値も本文中の値と全く同じになる。すなわち，両側5％水準で「有意差あり」と言ってよい。

（6）この問25において「特になし＋無回答（28.8%）」を欠損値とみなすと，2394×0.288＝689人が欠損値とみなされる。しかし，$n_{1.}$，$n_{.1}$に変化は無いので公式②で求めるp値の最小値は本文中の値と同じである。またN＝2394－689＝1705となり，670＋644＝1314＜1705より$n_{1.}+n_{.1}<N$なのでp値の最大値の計算には公式③が適用されるから，この値も本文中の値と同じになる。よって結論は本文同様「不明」になる。

（7）この問 48 において「特になし＋無回答（10.2%）」を欠損値とみなすと 2394×0.102＝244 人が欠損値とみなされるが，$n_{1.}$，$n_{.1}$に変化は無いので公式②で求める p 値の最小値は本文中の値と同じである。ただし，N＝2394－244＝2150 となり，1084＋1080＝2164＞2150 すなわち，$n_{1.}+n_{.1}>N$ となるから，本文中とは異なり，p 値の最大値の計算には公式④を使う。そこで $n_{0.}$＝2394－1084－244＝1066，$n_{.0}$＝2394－1080－244＝1070 を公式④に代入し，EXCEL の数式で計算すると $\max(p)$＝0.948247≒0.948 となり，本文中よりわずかに p 値の最大値が小さくなる。しかし，p 値の最小値は本文中と同じ 0.125 なので，p 値の最小値が α を超えることに変わりはない。よって，結論は本文中と同じく「有意差なし」である。

（8）証明は以下のとおり。欠損値を除く比率の差，すなわち $(n_{1.}-n_{.1})/N$が補遺 2 の表 8 のとき最大となることを証明する。表 5 の周辺度数を固定したとき A でn_{0m}の最大値，B でn_{m1}の最大値を表し，$n_{1m}=t$，$n_{mm}=s$，$n_{m0}=u$と表記すれば，一般に$n_{0m}=A-t-s$，$n_{m1}=B-u-s$と表現できる。すると証明すべきは $s=t=u=0$ のとき（表 8 のとき）欠損値を除く比率の差が最大となることである。$s=t=u=0$ のときの比率の差を D/N（0 以上 1 以下）と略記すると一般の場合の比率の差は $\{D-(s+t+u)\}/(N+s)$ と書ける。$D\geq D-(s+t+u)$ かつ $N\leq(N+s)$ より必ず $D/N\geq\{D-(s+t+u)\}/(N+s)$ である。よって表 8 のとき，欠損値を除く比率の差は最大となる。

（9）証明は以下のとおり。補遺 2 の表 9 または表 10 のときに欠損値を除く比率の差が最小となることを証明する。表 5 の周辺度数を固定したとき，A で n_{1m}の最大値を，B で n_{m0}の最大値を表し，$n_{0m}=t$，$n_{mm}=s$，$n_{m1}=u$ と表記すれば一般に $n_{1m}=A-t-s$，$n_{m0}=B-u-s$ と表現できる。すると証明すべきは $s=t=u=0$ のとき（表 9 または表 10 のとき）欠損値を除く比率の差が最小となることである。$s=t=u=0$ のときの比率の差を D/N とすると比率の差一般は $(D+s+t+u)/(N+s)$ となり，$(D/N)-\{(D+s+t+u)/(N+s)\}=\{D(N+s)-N(D+s+t+u)\}/\{N(N+s)\}=\{s(D-N)-N(t+u)\}/\{N(N+s)\}$。ここで $s(D-N)\leq0$ かつ $-N(t+u)\leq0$ かつ $\{N(N+s)\}>0$。よって，$(D/N)-[\{D+s+t+u)\}/(N+s)]\leq0$。したがって表 9 または表 10 のとき欠損値を除く比率の差は最小となる。

（10）1995 年 SSM 調査（社会階層と社会移動全国調査）の一環として行われたものである。

（11）具体的には層化 2 段確率比例抽出法（全国約 300 地点）が用いられている（原編 2000：付録 x vi 頁）。設計標本数は 1675 人で，そのうち回収標本数は 1214 人である。

参考文献

原純輔（編），2000，『日本の階層システム 1　近代化と社会階層』，東京大学出版会.

森口繁一・宇田川銈久・一松信，1987，『級数・フーリエ解析（岩波　数学公式Ⅱ）』（新装版），岩波書店.

NHK 放送文化研究所世論調査部（編），2008，『日本人の好きなもの　データで読む嗜好と価値観』，日本放送出版協会.

1995 年 SSM 調査研究会（編），1997，『1995 年 SSM 調査基礎集計表』，1995 年 SSM 調査研究会.

都築一治（編），1998，『職業評価の構造と職業威信スコア（1995 年 SSM 調査シリーズ 5）』，1995 年 SSM 調査研究会.

安田三郎，1969，『社会統計学』，丸善.

安田三郎・原純輔，1982，『社会調査ハンドブック〔第 3 版〕』，有斐閣.

（やまぐち　よう　現代社会学科）

2012 年 4 月 30 日受理